DOCUMENT RESUME

ED 389 739 TM 024 368

AUTHOR Stocking, Martha L.; Lewis, Charles
TITLE A New Method of Controlling Item Exposure in Computerized Adaptive Testing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-25
PUB DATE Aug 95
NOTE 34p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Selection; Test Construction; Test Format; *Testing Problems; *Test Items; Thinking Skills
IDENTIFIERS *Item Exposure (Tests); Paper and Pencil Tests; *Parallel Test Forms; Test Security

ABSTRACT

In the periodic testing environment associated with conventional paper-and-pencil tests, the frequency with which items are seen by test-takers is tightly controlled in advance of testing by policies that regulate both the reuse of test forms and the frequency with which candidates may take the test. In the continuous testing environment associated with more novel testing paradigms such as computerized adaptive testing (CAT), the computer itself can be used to control the frequency with which items are administered. This paper discusses previous methods of controlling item security in the continuous adaptive testing environment and presents a new method that overcomes some (but not all) of the disadvantages of previous attempts. The new method rethinks the use of the exposure control parameters in selecting each item to be administered. An extensive sample with this new method and a particular adaptive testing algorithm illustrates how concerns about test efficiency, parallelism, and security can be balanced. The target population was estimated using the method of R. J. Mislevy (1984) and a sample of over 5,000 real test-takers who took a linear 50-item paper-and-pencil analytical reasoning test to which the adaptive tests were designed to be parallel. (Contains 2 figures and 19 references.) (Author/SLD)

RESEARCH REPORT

A NEW METHOD OF CONTROLLING ITEM EXPOSURE IN COMPUTERIZED ADAPTIVE TESTING

Martha L. Stocking
Charles Lewis

Educational Testing Service
Princeton, New Jersey
August 1995

A NEW METHOD OF CONTROLLING ITEM EXPOSURE IN COMPUTERIZED ADAPTIVE TESTING

Martha L. Stocking
Charles Lewis

Educational Testing Service
Princeton, New Jersey 08541

July, 1995

A NEW METHOD OF CONTROLLING ITEM EXPOSURE IN COMPUTERIZED ADAPTIVE TESTING

Abstract

In the periodic testing environment associated with conventional paper-and-pencil tests, the frequency with which items are seen by test-takers is tightly controlled in advance of testing by policies that regulate both the reuse of test forms and the frequency with which candidates may retake the test. In the continuous testing environment associated with more novel testing paradigms such as computerized adaptive testing (CAT), the computer itself can be used to control the frequency with which items are administered. This paper discusses previous methods of controlling item security in the continuous adaptive testing environment and presents a new method that overcomes some (but not all) of the disadvantages of previous attempts. An extensive example with this new method and a particular adaptive testing algorithm illustrates how concerns about test efficiency, parallelism, and security can be balanced.

Key Words: computerized adaptive testing, item exposure control, test security, exposure rates.

A NEW METHOD OF CONTROLLING ITEM EXPOSURE IN COMPUTERIZED ADAPTIVE TESTING

Introduction

Every year millions of conventional paper-and-pencil tests are administered by various national testing agencies. These tests are typically "high stakes" tests in that important decisions about test-takers are based, in part, on test scores. In secure conventional paper-and-pencil testing, large numbers of candidates take the same or parallel linear test forms at a few fixed administration dates scheduled throughout some time period. By "secure" we mean that a great deal of time and effort is spent by test agencies to insure that no test-taker has access to test questions in advance of test administration. In this context, the frequency with which a single item might be seen by a single test-taker can be tightly controlled in advance of testing through policies that regulate both the reuse of test forms and the frequency with which candidates may retake the test. Such a system of test administration and its associated policies may be called periodic testing.

Adaptive tests are tests in which items are selected from a large pool of items to be appropriate for a test-taker (the test "adapts" to the test-taker). All but a few proposed designs have assumed that items would be chosen and administered to test-takers on a computer, hence the term "computerized adaptive testing" or CAT. (See Lord (1980) or Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen (1990) for a more detailed description of adaptive testing.) In an environment where tests are computer administered, it is a natural extension to utilize the computer for such administrative activities as scheduling, score reporting, protecting item security, and so forth. In this environment continuous, as opposed to periodic, testing becomes possible.

It is, of course, also possible to conceive of conventional paper-and-pencil testing in a continuous testing environment, although the authors know of no high stakes paper-and-pencil linear conventional tests administered in this fashion. It is quite likely that the security problems of such an administrative mode are difficult to overcome for reasonable cost. Likewise, CAT is by no means the only type of test that is convenient to administer via computers. (See for example, Sheehan & Lewis (1992) on computerized mastery testing.) It is likely that different types of high-stakes continuous computer administered tests have different kinds of security problems.

In this paper we discuss issues of item security only in the context of CAT. In the next section we briefly provide more details about CAT and a particular adaptive testing algorithm that we will use in a later example. The following section provides a history of methods, with advantages and disadvantages, that seek to provide secure testing in the continuous testing environment of CAT. In the subsequent section, a new method of controlling item exposure in CAT is presented, followed by an example using this new method.

The Weighted Deviations Model

As noted by Davey & Parshall (1995) high-stakes adaptive testing has at least three goals: 1) to maximize test efficiency by selecting the most appropriate items for a test-taker, 2) to assure that the tests measure the same composite of multiple traits for each test-taker by controlling the nonstatistical nature of items included in the test, and 3) to protect the security of the item pool by controlling the rates at which items can be administered. These goals often compete with one another.

Different approaches to each of these goals yield different algorithms for adaptive testing. The particular algorithm used in this paper is the Weighted Deviations Model (WDM) developed by Stocking & Swanson (1993) and Swanson & Stocking (1993). This paradigm is characterized by flexible approaches to all three goals of adaptive testing.

In general, any CAT algorithm implicitly orders the items in the pool in terms of their desirability for selection as the next item. Differences in ordering typically reflect particular definitions of item optimality and particular methods of estimating ability. Any attempt to control the exposure of items can then be viewed as modifications imposed on this ordering.

In the WDM the item pool is ordered by employing a methodology from the decision sciences that models the behavior of expert test specialists. The WDM ordering explicitly takes into account nonstatistical item properties or features along with the statistical properties of items. This is to insure that each adaptive test produced from a pool matches a set of test specifications and is therefore as parallel as possible to any other test in terms of content and type of items, while being tailored to an individual test-taker in terms of appropriateness. The desired balance between measurement and construct concerns is reflected by the weights given to them, which are chosen by the test designer. The WDM approach also allows specification of overlapping items that may not be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or any other feature that test specialists deem important. Thus at each item selection in the WDM, the pool or an appropriate subset of the pool is ordered from most desirable (smallest weighted deviations from

desirable test properties) to least desirable (largest weighted deviations from desirable test properties).

In summary, in the WDM, the next item selected for administration is the item that simultaneously

1) is the most appropriate possible at a test-taker's estimated ability level, while

2) contributing as much as possible to the satisfaction of all other constraints.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the previous item was in a block), starts a new block, or is in no block.

In the particular version of the WDM used in this paper, the measure of the appropriateness of the item is the Fisher item information function (Lord, 1980, equation 5-9) and the estimate of ability is maximum likelihood (Lord, 1980, equation 4-31), although other measures of the statistical properties of items (see for example, Chang, 1995) and other estimates of ability (see for example, Davey & Parshall, 1995) are possible.

Previous Methods of Controlling Item Exposure

Any scheme that seeks to control the exposure of items employs mechanisms that override the optimal ordering of items, thus degrading the quality of the adaptive test. Longer tests are therefore required to achieve the level of psychometric efficiency obtained when no exposure control is exercised, but longer tests may be viewed as a reasonable exchange for greater

item and test security. In this section, we describe a number of exposure control methods that have been tried out in the past. The discussion proceeds in order of the complexity of the scheme.

Simple Randomization

Early theoretical investigations of CAT ignored the problem of item exposure (see, for example, Lord, 1970). Procedures that seek to prevent the overexposure of initial items developed when the prospects of actual implementation became more certain. Lord (1977), McBride & Martin (1983), Stocking (1987), and Weiss (1978) implemented strategies typical of these first attempts. In this approach, the selection of the next item to administer is no longer based solely on the evaluation of items for optimality at the current ability estimate, however optimality may be defined in a particular application. Rather, a group of items is identified that are roughly equal in optimality and the next item is chosen randomly from this group.

In the context of the WDM, the first item to be administered would be chosen randomly from, say, the top five items in the list of items ordered as described above. The second would be selected randomly from a group of four most desirable items, the third item from a group of three, the fourth randomly from a group of two and the fifth and subsequent items chosen to be optimal. The assumption underlying this approach is that after some number of initial items, test-takers will be sufficiently differentiated so that subsequent items will vary a great deal.

Many variations on this theme are possible, including the possibility of never choosing the next item optimally with certainty, that is, the minimum group size is always two or greater. This latter approach recognizes that in

spite of randomization on initial items, test-takers with similar abilities may receive many of the same items subsequently unless attempts are made to control the exposure of items later in the test.

The advantage to these kinds of schemes is that they are simple to implement and easily understood. They can be elaborated to make the random selection process depend upon the estimated ability of the test-taker by specifying different group sizes for different levels of estimated ability. However, the success of such schemes is difficult to predict with complex but realistic item pool structures and test specifications, and may not prevent overuse of some items (for an example of this, see Stocking (1993), Figure 1). Moreover, it is difficult to determine the best sequence of group sizes from which the random selection is made by anything other than time-consuming trial and error with no certainty of success and with no easy generalizability to different item pool and test structures.

Thomasson & Drasgow (1990)

This procedure, called the "INFO4" procedure, is described in Segall (1994). In the application described, at every item selection the items in the entire pool are ordered from highest to lowest based on their Fisher information at the current level of estimated ability. These values are then raised to the fourth power, an ad hoc decision made in order to emphasize differences in information (Drasgow, personal communication, January 11, 1995) according to Tukey's ladder of re-expressions (Mosteller & Tukey, 1977). A maximum is placed on these values, the values are then normed to sum to one and a cumulative function is formed. A random number is generated and the location of the corresponding item is found for the value of the random number, interpreted as a cumulative probability. This item then becomes the

next item to be administered. If such a method were used with the WDM, the pool would be ordered not by information, but rather by the desirability criteria incorporating both statistical and nonstatistical item features as described previously.

The INFO4 procedure avoids the problem of determining the best sequence of group sizes that characterizes the simple randomization method. It is similar to the simple randomization approach with randomization at every item selection. Also, intrinsic to the INFO4 method is the implicit dependence of the randomization on the current estimated ability level. However, this scheme does not prevent high exposure rates for some items, as reported in Segall (1994, Table 2.1). In addition, no investigations have been conducted of different transformations of the information function, and it is likely that this procedure depends on the nature of the particular item pool for which it was developed (Drasgow, personal communication, January 11, 1995). It may be difficult or impossible to generalize to other pools.

Sympson & Hetter (1985)

The two procedures described above attempt to increase item security by indirectly reducing item exposure. Sympson & Hetter (1985) tackle the issue of controlling item exposure directly in a probabilistic fashion.

This procedure considers a test-taker randomly sampled from a typical group of test-takers and distinguishes between the probability P(S) that an items is selected as the best next item by some CAT algorithm and P(A|S), the probability that an item is administered, given that it has been selected. The procedure seeks to control the overall probability that an item is administered, P(A), where P(A) = P(A|S) * P(S), and to insure that the maximum value over all P(A)s is less than some value r. This r is the desired (not

observed) maximum rate of item usage.

The 'exposure control parameters', P(A|S), one for each item, are determined through a series of simulations (iterative adjustment simulations) using an already established adaptive test design and simulees drawn from a typical distribution of ability. Following each simulation, the proportion of times each item is selected as the best item, P(S), and the proportion of times each item is administered, P(A), are separately tallied. If P(S) is less than or equal to r, then P(A|S) is set to one for the next iteration, insuring that $P(A) = P(A|S)*P(S) \leq r$. If P(S) is greater than r, then P(A|S) is set to r/P(s) for the next iteration, again insuring that $P(A) \leq r$. The simulations continue until the P(A|S) have stabilized and the maximum observed P(A) for all items is approximately equal to the desired value of r. Note that there is no guarantee that this procedure will eventually stabilize, and indeed, it may not (see below).

Once the exposure control parameters have been established (as well as during the iterative adjustment simulations), they are used in item selection as follows:

1) Select the next item for administration.
2) Generate a random number uniformly distributed between zero and one.
3) If the random number is less than or equal to the exposure control parameter for the selected item, administer the item.
4) If the random number is greater than the exposure control parameter for the optimal item, do not administer the item and remove it form the pool of remaining items for this test-taker. Repeat this procedure for the next-most-optimal item. Continue until an item

is administered.

If the adaptive test is of length n, then there must be at least n items in the pool with exposure control parameters of one. If there were not, then for some test-takers there might not be enough items in the pool to administer a complete adaptive test. In the case where there are not n such items, Sympson & Hetter suggest the reasonable procedure of sorting the values of the exposure control parameters and setting the n largest to one. This has the effect of increasing the exposure rate for the items that are least popular -- a conservative approach.

Stocking (1993) extended the Sympson & Hetter approach to item pools with complex structures and adaptive tests with complex test specifications. In these extensions, the basic procedure is applied to blocks of items as well as to stimulus material, which, in general, will have different exposure rates than items associated with stimulus material.

The advantage of the extended Sympson & Hetter approach is that one obtains direct control of the probability that an item is administered, P(A), in a typical population of test-takers. However, the simulations required to obtain estimates of the P(A|S) for each item are time-consuming for pools and test specifications with complex structures. If an item pool is changed, even by the addition or deletion of a single item, or if the target population changes significantly, the adjustment simulations must be repeated.

Moreover, if the structure of the item pool is not a good match with the structure of the test specifications, it is possible for the extended Sympson & Hetter procedure to diverge, that is, it may not be possible to obtain stable estimates of the P(A|S) for each element in the pool (see Stocking, 1993, Figure 4, for an example). This happens because of the 'fixup' to

insure complete adaptive tests -- setting the n highest P(A|S) to one. This fixup seems to work well if all n of the high P(A|S)s are not too different from one. However, if some are very different, this fixup can cause thrashing by repeatedly setting low P(A|S)s back to one or alternating among several items with low P(A|S)s. This prevents smooth convergence of the procedure. A solution to this problem is to construct a context in which there is a better match between pool structure and test specifications, either by enriching the pool or by simplifying test structure. Either of these may be difficult to accomplish.

A New Method of Controlling Item Exposure

The new method of controlling item exposure seeks to overcome the lack of robustness of the extended Sympson & Hetter procedure by rethinking the use of the exposure control parameters P(A|S) in selecting each item to be administered. It can also be viewed as a cousin of the INFO4 method of Thomasson & Drasgow in that some function is formed and treated as a cumulative probability for the purpose of selecting an item.

A helpful way of viewing this new procedure is that it formally models the Sympson & Hetter procedure, in the following fashion: Consider the list of items ordered by the WDM model from most desirable to least desirable, and the associated $P_i(A|S)$. (In what follows, we will abbreviate this notation to P_i for convenience.) The operant conditional probabilities of administration for each item, k_i, are not the simple P_i but rather as follows:

$k_1 = P_1$, the probability that item 1 is administered given that it is selected,

$k_2 = (1-P_1) * P_2$, the probability that item 1 is rejected given that it

is selected and the probability that item 2 is administered given that it is selected,

$k_3 = (1-P_1) * (1-P_2) * P_3$, the probability that the first two items are rejected given selection and that item 3 is administered given that it is selected,

and so forth.

The sum S of these probabilities must equal one for some event to occur, that is, some item to be administered. If they do not sum to one, it may occur that no item will be administered. If S is not one, we can define adjusted probabilities whose sum is guaranteed to equal one by dividing each k_i by the sum S. This adjustment of probabilities is the analog of the fixup recommended in the extended Sympson & Hetter procedure in that it guarantees that an item (and therefore a complete adaptive test) can always be found for administration. However, if the list of items is long the adjustment to individual operant probabilities may be quite small, thus increasing the chance for smooth convergence of the procedure.

The distribution of adjusted probabilities is now a multinomial distribution and we want to sample one event from it, that is, we want to administer an item. To do this, we form the cumulative distribution, generate a random number between zero and one, and locate the item to be administered. We eliminate all items appearing in the ordering before the item to be administered from further consideration for this adaptive test. This elimination of items accords with the definition of the operant probabilities given above in that the operant probability of selecting item i includes the probabilities of rejecting all items before item i in the ordered list.

If the list of elements ordered by the WDM model contains both discrete

items and stimuli with associated items, then the process of selecting an event from the ordered list is modified slightly. The P_i for a set of items is considered to be the P_i for the stimulus of the set, and the computation of the k_i for all possible events, now a mixture of discrete items and stimuli, proceeds as before. If a stimulus is the element sampled from the cumulative multinomial distribution, all elements preceding the sampled stimulus are eliminated from further consideration, and a new ordered list is then prepared containing only items associated with the stimulus. The sampling from this second multinomial distribution proceeds as with discrete items, but is restricted to items within the set.

The most desirable items in any selection of an item tend to have not only low deviations, but also low P(A|S) since items with low deviations tend to be popular items. In some circumstances it may be desirable to move towards using these items with low deviations, even at the expense of their over exposure. This type of control can be provided to the test designer by the use of an exponent on the probabilities of rejection of items that reflects the magnitude of the weighted deviation relative to the previous element in the ordered list. That is, each $(1-P_i)$ can be raised to the power $(1+C*\Delta_i)$ when forming the operant probabilities k_i, where C is some constant set by the test designer and Δ_i is the weighted deviation of the item relative to the previous item in the list.

Table 1 gives an example that shows the effect of this procedure. The top part of the table gives information about five hypothetical items. The weighted deviations and the relative weighted deviations are listed in the second and third columns, respectively. The probabilities of administration and rejection of each item, given the item is selected are in the remaining

two columns. The next three parts of Table 1 show the effect on the cumulative multinomial distribution when the coefficients used in the exponents of the probabilities of rejection are 0, .5, and 1.0 respectively.

The entries for C = 1, for example, are computed as follows:

$k_1 = .10$

$k_2 = (.9^{(1+6)})\ (.14) = .07$

$k_3 = (.9^{(1+6)})\ (.86^{(1+1)})\ (.14) = .05$

$k_4 = (.9^{(1+6)})\ (.86^{(1+1)})\ (.86^{(1+0)})\ (.16) = .05$

$k_5 = (.9^{(1+6)})\ (.86^{(1+1)})\ (.86^{(1+0)})\ (.84^{(1+1)})\ (.22) = .05.$

The sum of these five operant probabilities is .32. The adjusted probabilities are obtained by dividing each operant probability by this sum, that is, .10/.32, .07/.32, .05/.32, and so forth. The cumulative probabilities are obtained by successive addition from the adjusted operant probabilities.

The probability of selecting the most desirable item, which also has the lowest exposure control parameter, rises from .18 when relative weighted deviations are not emphasized further (C = 0), to .24 when they are emphasized a moderate amount (C = .5), to .31 when they are emphasized more heavily (C = 1). By increasing the coefficient C, the test designer can increase the influence of the weighted deviations on the selection of each item at the expense of increasing the exposure of desirable items.

Table 1: The Effect of the Exponent on Probabilities of Rejection in the Computation of Operant Probabilities

Item	Weighted Deviations	Relative Deviations (Δ)	P(A|S)	1 - P(A|S)
1	0	—	.10	.90
2	6	6	.14	.86
3	7	1	.14	.86
4	7	0	.16	.84
5	8	1	.22	.78

C = 0

Item	k_i	Adjusted	Cumulative
1	.10	.18	.18
2	.13	.23	.41
3	.11	.19	.60
4	.11	.19	.79
5	.12	.21	1.00

C = .5

Item	k_i	Adjusted	Cumulative
1	.10	.24	.24
2	.09	.22	.46
3	.07	.17	.63
4	.07	.17	.80
5	.08	.20	1.00

C = 1

Item	k_i	Adjusted	Cumulative
1	.10	.31	.31
2	.07	.22	.53
3	.05	.16	.69
4	.05	.16	.84
5	.05	.16	1.00

The adjustment of the sum of the operant probabilities to one, as well as the addition of the exponent on the probabilities of rejection, move the new procedure away from a strict modeling of the Sympson & Hetter procedure. That is, nominal operant probabilities, k_i, are no longer equal to actual

operant probabilities because of these modifications.

In summary, the new procedure consists of the following steps:

1) Choose a value of C, to reflect the relative importance of the deviations from desirable measurement and construct properties of the adaptive test when compared to item security.

2) Establish the exposure control parameters for each item in the same manner as done in Sympson & Hetter iterative adjustment simulations.

3) When selecting the next item in each simulation (and with the final exposure control parameters once they have been established):

   a) Form a list of items ordered by their desirability.

   b) For each element i in the list, form the operant probabilities k_i, where

$$k_i=\{\prod_{j=1}^{i-1}(1-P_j^{(1+C*\Delta_j)})\}*P_i \ .$$

   c) If necessary, adjust the operant probabilities so that they sum to one by dividing each value by their sum.

   d) Form the cumulative distribution. Generate a random number uniformly distributed between zero and one.

   e) Find the corresponding element in the cumulative distribution.

   f) Remove all elements preceding the one selected from further consideration in this adaptive test.

   g) If the element selected is a stimulus for a set of items, repeat steps a) through e) for items belonging to this set.

This procedure provides a smaller adjustment to the P(A|S) than the Sympson & Hetter procedure in order to guarantee that a complete adaptive test can always be found. The iterative adjustment simulations to determine the P(A|S) are therefore more likely to converge smoothly to values that appropriately reflect the intended population of test-takers. At the same time this new procedure retains the advantage of the Sympson & Hetter procedure in that it provides direct control over P(A) for each item when adaptive tests are drawn for administration to the intended population.

While offering these advantages over the Sympson & Hetter procedure, this new procedure retains the major disadvantages of Sympson & Hetter. The iterative adjustment simulations are time-consuming for pools and test specifications with complex structures. And if an item pool is changed or if the target population changes significantly, the iterative adjustment simulations must be repeated.

An Example

In this section we present an extensive examination of the trade-offs that can be made between the measurement, content, and security properties of adaptive tests drawn from a particular item pool for a particular target population. The target population was estimated using the method of Mislevy (1984) and a sample of over 5000 real test-takers who took a linear 50-item paper-and-pencil analytical reasoning test to which the adaptive tests are designed to be parallel.

The Item Pool

Available to the authors was a large pool of items and sets of items measuring various aspects of Analytical Reasoning. There were a total of 660

elements in this pool -- 578 items and 82 stimuli. Of the 578 items, 491 were associated with the 82 stimuli and the remaining 87 items were discrete items. The items were calibrated on large samples (2000+) of test-takers using the 3-parameter logist item response model and the computer program LOGIST (Wingersky, 1983), and placed on the same IRT metric using the transformation methodology of Stocking & Lord (1983). The mean item discrimination was .75 with a standard deviation of .25; the mean item difficulty was .06 with a standard deviation of 1.39; and the mean pseudo-guessing parameter was .16 with a standard deviation of .10.

The Adaptive Tests

Items were drawn from this pool using the WDM to form (fixed length) adaptive tests of 35 items, subject to 34 constraints on their content. Stimuli were drawn subject to 9 constraints on the nature of the stimuli. Thus a total of 43 constraints controlled the nonstatistical features of items and stimuli appearing in an adaptive test. These constraints had relative weights that varied from 20.0, indicating that it was very important for an adaptive test to have items and/or stimuli with these features, to 1.0, indicating that it was substantially less important for an adaptive test have these features. The importance of measurement appropriate for a test-taker was reflected in the weighting of item information (Lord, 1980, equation 5-9) at 20.0.

In addition to this relatively complex test structure, item selection was further restricted by the specification of 75 overlap groups. Items and stimuli belonging to an overlap group may not appear in the same adaptive test with other items and stimuli appearing in the same overlap group. When a stimulus appears in an overlap group, all items associated with that stimulus

are included by implication. There were a total of 312 entries in the 75 overlap groups.

The Simulations

Ten iterative simulations were performed for each of four target maximum desired exposure rates (r = .1, .2, .3, and .4) combined with each of three values of C in the exponent of the conditional probabilities of rejection (.0, .5, and 1.0) for a total of 120 (4 x 3 x 10) iterative simulations. Each simulation was performed with 1170 simulated examinees (simulees); Sympson & Hetter recommend sample sizes of at least 1000. Within each of the 12 sequences of ten iterative simulations, exposure control parameters started with values of 1.0, as recommended by Sympson & Hetter, and were adjusted between simulations as suggested by Sympson & Hetter. Within each of the 120 simulations, item selection was performed using the new (multinomial) procedure described above. All 12 sequences of iterative simulations converged to stable estimates of the exposure control parameters and maximum probabilities of administration.

The Results

The results of each of the 12 final iterative simulations were incorporated into single number summaries for convenience (in displaying the results) as follows:

1) Measurement efficiency

Although complete conditional standard error of measurement curves were available, in the context of this study it was easier to interpret the results for the target population of interest by incorporating this information into an estimate of adaptive test reliability as suggested by Green, Bock, Humphreys, Linn, &

Reckase (1984, equation 6).

2) Measurement of the intended construct

Information was available concerning the extent of violations for each nonstatistical constraint on item selection (that is, we exclude the information constraint). These data were weighted by the relative weight assigned to each constraint and them summed over the 1170 simulees to give the total weighted deviations for a typical group of size 1170 drawn from the target population.

3) Item security

The largest observed probability of administration for any element in the pool for a sample of size 1170 drawn from the target population was used as a single number summary of item security. To the extent that the estimate of the target population matches the true population of test-takers, this maximum observed probability from the simulations should agree with data from real test-takers.

Figure 1 displays one method of analyzing the results of this experiment. In this figure, the horizontal axis is the targeted (not observed) value of the maximum exposure rate, r, of any element in the pool, and has values of .1, .2, .3, and .4. The vertical axis is the $\log_{10}$ (chosen to improve the readability of the Figure) of the total weighted deviations for a sample of size 1170. Lines are drawn connecting the values of the total weighted deviations for a particular value of C (0, .5, or 1.0) at the different values of target maximum exposure rate. The numbers appearing beside each point indicate the resultant adaptive test reliability.

A single square point is also drawn on the figure, at r = .4. This

represents the total weighted deviations obtained if there were no control on the exposure of items (target maximum r of 1.0). In this case, the value of C can be anything because the most desirable item is always picked. This represents a lower bound on how much the total weighted deviations can be reduced for this pool.

At each level of target maximum exposure, increasing C reduces the total weighted deviations, although the difference when item security is less of a concern (r of .3 and .4) is small. When item security is more of a concern (r of .1 or .2) the largest effect is gained by moving C from 0. to .5, with less of an effect when moving from .5 to 1.0. Increasing C also increases test reliability, as expected, since it improves the selection of items statistically appropriate for simulees. However, this difference appears quite small.

Figure 1 tells only two-thirds of the story when considering trade-offs between efficiency, content, and security. What is absent is information about the observed (not targeted) level of security actually obtained. This information is shown is Figure 2, which is more difficult to interpret, but also more informative.

In Figure 2, the horizontal axis is the observed maximum probability of administration across the entire item pool from the final iterative simulation for each condition. This horizontal axis now goes from zero to one. The three lines plotted are for different values of C, as in Figure 1, and are plotted with the same plotting symbols in both figures. The numbers next to each plotted point are now the targeted maximum probability of administration.

If the targeted maximum probability of administration is .1, increasing C reduces the total weighted deviations, but at the expense of increasing the

observed maximum exposure rate from close to .1 to over .2. If the targeted maximum probability of administration is .2, increasing C reduces the total weighted deviations substantially, but at the expense of increasing the observed maximum exposure rate from close to .2 to over .5. If the targeted maximum exposure rate is higher than .2, increasing C does not lower the total weighted deviations very much, but increases the observed maximum exposure rate substantially.

For this pool, for this population, for the WDM item selection method, and for the multinomial method of exposure control, C = 0 gives observed maximum exposure rates closest to target maximum exposure rates. In this condition, substantial reduction in total weighted deviations can be obtained simply by increasing the targeted maximum exposure rate. These reductions are most striking when the target is increased from .1 to .2 and .2 to .3. If reducing the total weighted deviations is very important, then a target maximum exposure rate of .3 appears satisfactory. However, if item and pool security is of primary importance, and the difference in total weighted deviations does not represent the measurement of substantially different constructs, a target maximum exposure rate of .2 may offer a good compromise.

Discussion

The continuous testing environment requires renewed attention to item and test security concerns that have previously been resolved through administrative procedures. Because of the size and cost of adaptive test item pools, it is unlikely that it will be possible to have more than a few item pools in operational use at the same time. In a more realistic approach, the protection of item security assumes the form of suboptimal item selection

within a single item pool to decrease the frequency of use of the best items in the pool.

One simple previously published method for controlling the frequency of item administration sought to accomplish this by randomly selecting an item for administration from a group of items of approximate equivalent optimality. In order to avoid issues of optimum group sizes, a second previously published method formed a function of item parameters, treated this function as a cumulative probability, and selected from this cumulative probability at each item selection. These methods controlled exposure rates only indirectly and did not solve the problem of high exposure rates for some items.

The Sympson & Hetter approach develops an exposure control parameter for each item that directly controls the frequency of administration for that item in reference to a particular distribution of test-taker ability. This approach, however, is not adequate for item pools with complex structures and adaptive tests with complex specifications, particularly when specifications do not conform well with the structure of the pool. In some cases, the Sympson & Hetter approach can fail to converge entirely (see Stocking, 1993, for examples).

In this paper we propose a new method of imposing exposure control on the selection of the next item. In many respects this new method can be viewed as modeling the Sympson & Hetter procedure by forming a cumulative multinomial distribution from the true operant probabilities of administration given selection for each element in a list of elements ordered by desirability. In the particular example, the WDM model was used to form the ordered list, but in theory, any method of ordering items can be used. This cumulative multinomial distribution may need to be adjusted so that the

sampling of one event (the selection of an item) is guaranteed, but this adjustment is apt to be smaller than that used in the Sympson & Hetter procedure and is therefore less likely to cause convergence problems. The method can be easily extended to include sets of items as well as discrete items. The method also allows test designers to specifically control the trade-offs between the efficiency and parallelism of adaptive tests and the need for security, as shown in the example.

However, this new method retains three of the major disadvantages of the Sympson & Hetter procedure. First, the process of iterative simulations with adjustments can be tedious and time-consuming. Second, the exposure control parameters computed by the procedure are dependent upon a specific pool of items and test structure. If the pool is augmented or reduced, or the test structure changed significantly, then exposure control parameters must be redeveloped. Finally, the exposure control parameters are in reference to a specific target estimated distribution of true ability. To the extent that this distribution does not accurately reflect the true distribution, or if the reflection is accurate but the distribution changes over time, for example, exposure control parameters must be redeveloped.

It seems clear that the new multinomial exposure control procedure as well as the Sympson & Hetter procedure from which it is derived do not have all the features one might eventually require in operational secure continuous adaptive testing. For example, although the overall exposure rate of an item is controlled, its exposure conditional on ability is not. Thus an item may be exposed to nearly all test-takers at a particular ability, even though its overall exposure rate is low. If this is identified as a problem, it may be necessary to develop new methods that control conditional exposure. Moreover,

although exposure rate is controlled across a distribution of ability, it is not controlled across candidate volume. An item with an exposure rate of .1 will only be seen by approximately 10% of test-takers, but if there are a million test-takers, the absolute exposure will be quite high. Further research clearly remains to be done in this area if continuous adaptive testing is to become a secure alternative to periodic conventional paper-and-pencil testing.

References

Chang, H. (1995). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. April, 1995, San Francisco, CA.

Davey, T., and Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association. April, 1995, San Francisco, CA.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer assisted instruction, testing, and guidance. New York: Harper and Row.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McBride, J. R., and Martin, J. T. (1983) Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.) New Horizons in Testing (223-236). New York: Academic Press.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mosteller, F., and Tukey, J. W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Segall, D. O. (1994). CAT-GATB simulations studies. San Diego, CA: Navy Personnel Research and Development Center.

Sheehan, K., and Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. Applied Psychological Measurement, 16(1), 65-76.

Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm. (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7(2), 201-120.

Stocking, M. L., and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Swanson, L., and Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-166.

Sympson, J. B., and Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing, as described in Wainer, et al., (1990).

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J.,Steinberg, L., and Thissen, D. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (Ed.) (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Wingersky, M. S. (1983) LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.),

Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

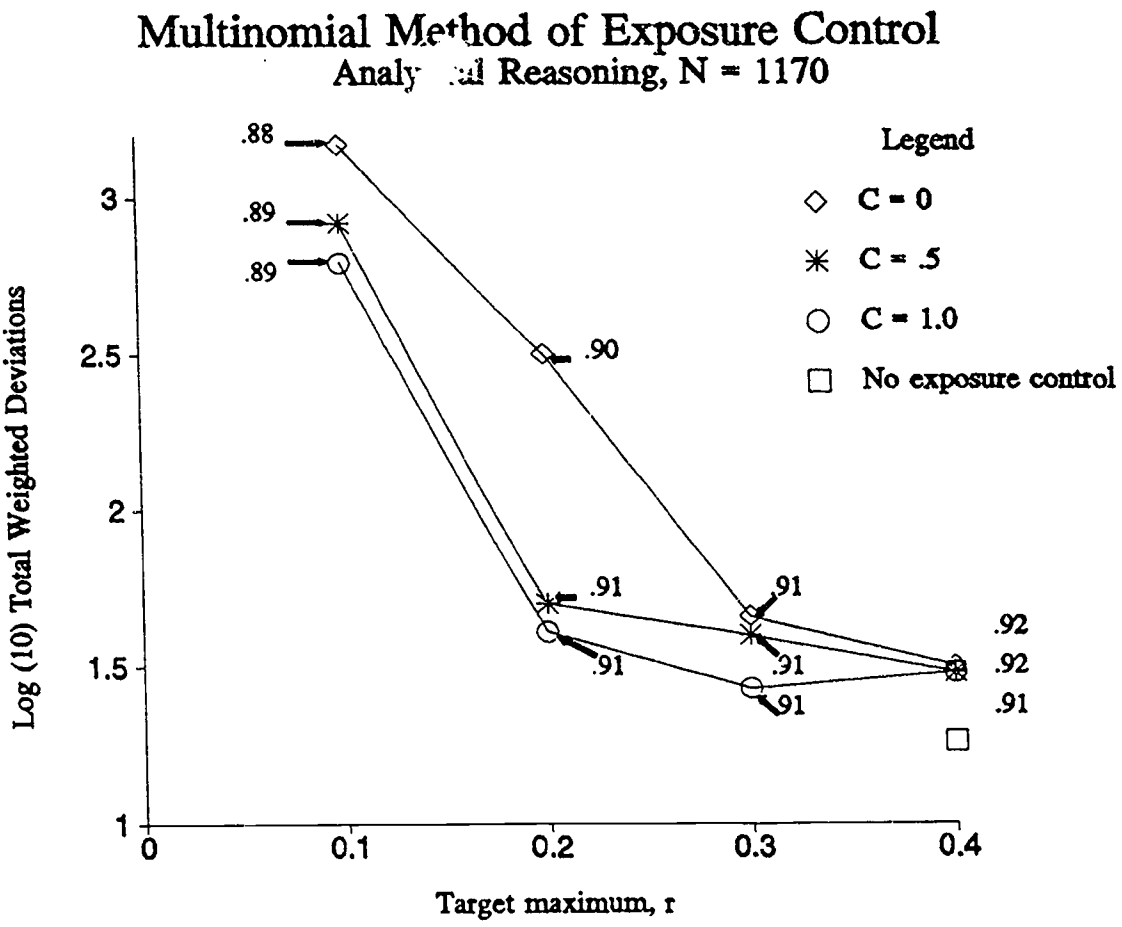

Figure 1: The effect of different values of C on total weighted deviations and reliability.

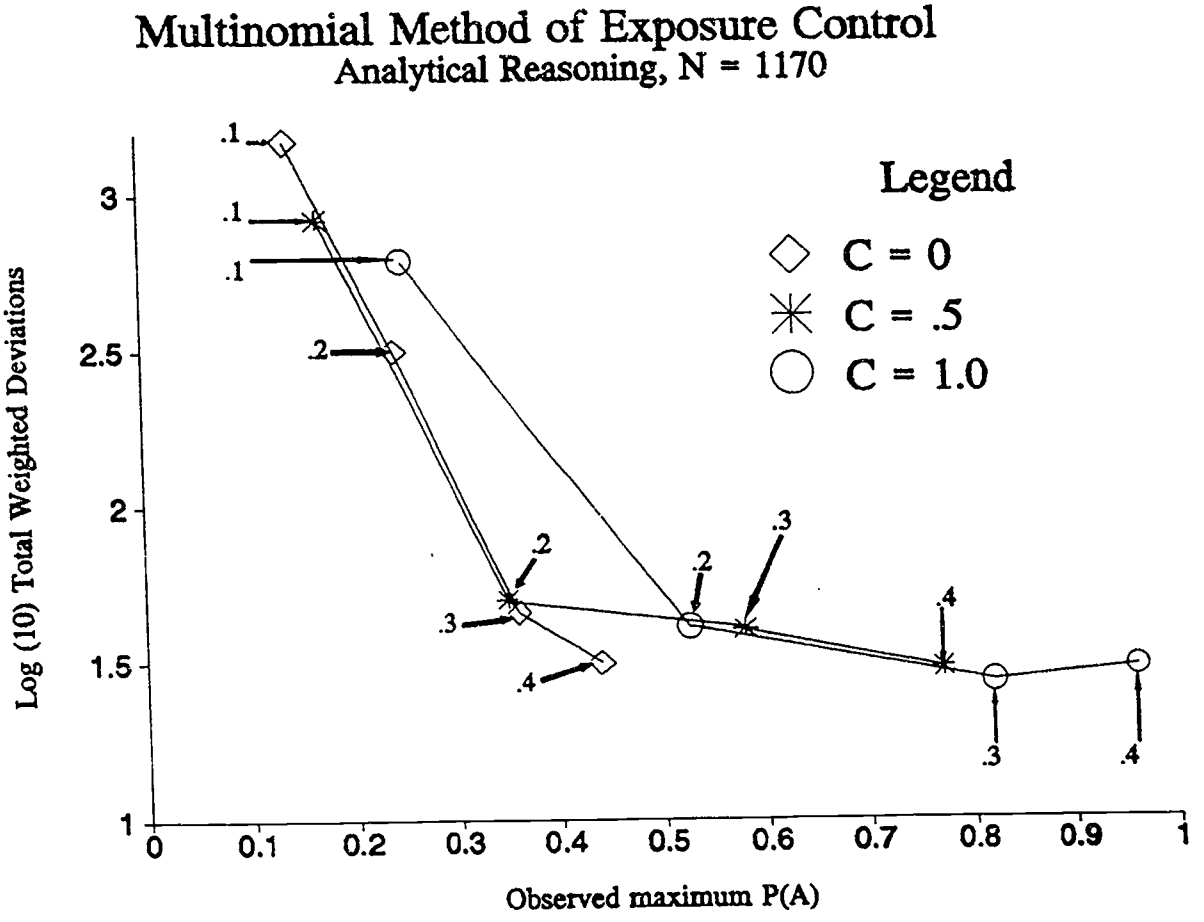

Figure 2: The effect of different values of C on total weighted deviations and observed maximum exposure rates.